University of Alberta

Improved Crystal Structure Refinement Through Maximum Likelihood

by

Navraj Singh Pannu

A thesis submitted to the Faculty of Graduate Studies and Research in partial fulfillment of the requirements for the degree of Master of Science

in

Applied Mathematics

Department of Mathematical Sciences

Edmonton, Alberta

Spring 1998

National Library of Canada
Acquisitions and Bibliographic Services
395 Wellington Street
Ottawa ON K1A 0N4
Canada

Bibliothèque nationale du Canada
Acquisitions et services bibliographiques
395, rue Wellington
Ottawa ON K1A 0N4
Canada

Your file Votre référence

Our file Notre référence

The author has granted a non-exclusive licence allowing the National Library of Canada to reproduce, loan, distribute or sell copies of this thesis in microform, paper or electronic formats.

The author retains ownership of the copyright in this thesis. Neither the thesis nor substantial extracts from it may be printed or otherwise reproduced without the author's permission.

L'auteur a accordé une licence non exclusive permettant à la Bibliothèque nationale du Canada de reproduire, prêter, distribuer ou vendre des copies de cette thèse sous la forme de microfiche/film, de reproduction sur papier ou sur format électronique.

L'auteur conserve la propriété du droit d'auteur qui protège cette thèse. Ni la thèse ni des extraits substantiels de celle-ci ne doivent être imprimés ou autrement reproduits sans son autorisation.

0-612-28974-5

Canada

University of Alberta

Faculty of Graduate Studies and Research

The undersigned certify that they have read, and recommend to the Faculty of Graduate Studies and Research for acceptance, a thesis entitled Improved Crystal Structure Refinement Through Maximum Likelihood submitted by Navraj Singh Pannu in partial fullfillment of the requirements for the degree of Master of Science in Applied Mathematics.

Byron Schmuland (supervisor)

Randy J. Read (co-supervisor)

Michael N. G. James

Douglas P. Wiens

22 Dec 97

Date

To jojo

Abstract

To elucidate the mechanism of a biological process, structural information of the molecules involved is often necessary. An important method for the determination of a molecule's three dimensional structure is X-ray Crystallography. In order to obtain the most accurate atomic coordinates, structural refinement is an essential part of a crystal structure determination. The refinement of crystal structures is commonly based on least-squares methods. However, these procedures are not optimal, since conditions necessary for the application of a least-squares target are not satisfied. Therefore, a more general maximum likelihood analysis is considered and three maximum likelihood targets have been implemented in the refinement packages CNS, TNT and X-PLOR. Preliminary tests with protein structures give dramatic results. Compared to least-squares refinement, maximum likelihood refinement can achieve more than twice the improvement in average phase error. With the inclusion of experimental phase information, a maximum likelihood strategy can further improve a model over least-squares.

Acknowledgements

This work would not have been completed without the guidance, support and assistance of my two supervisors, Dr. Randy Read and Dr. Byron Schmuland. I have enjoyed my two years as an undergraduate and two years in graduate studies under the supervision of Dr. Read in which this work was completed. I have found these four years challenging and thought provoking. Dr. Read provided the perfect atmosphere for this work to take place as well as giving me the freedom to explore new possibilities. Dr. Read's derivation and implementation of the likelihood function MLF1 is discussed in Chapter 2. Furthermore, the cross-validated σ_A estimation, for which the success of the discussed likelihood function is crucial, was developed and implemented by Dr. Read.

I am grateful for Dr. Schmuland for keeping me focused during my two years of graduate studies. Dr. Schmuland provided many useful discussions to clarify the ideas expressed here.

I wish to thank my defence committee members Dr. Michael James and Dr. Douglas Wiens who further clarified these ideas.

Financial support for this work was provided by the Natural Sciences and Engineering Research Council of Canada, the Alberta Heritage Foundation for Medical Research and a Walter H. Johns Scholarship. Financial support for the presentation of this work at conferences was provided by the Collaborative Computing Project 4, the European Union, the American Crystallographic Association, the Pittsburgh Supercomputer Center, the International Union of Crystallography Computing Commission and a J. Gordin Kaplan Scholarship.

Many individuals helped in the implementation and programming of the likelihood functions in the different refinement packages. Dr. Bart Hazes and Mr. Steven Ness

helped greatly in the incorporation of the likelihood functions into X-PLOR. Dr. Dale Tronrud, Dr. Marie Fraser and Dr. Anita Sielecki provided many useful discussions for the implementation of the likelihood functions into TNT.

Dr. Rik Wierenga kindly provided data and structures for the gTIM test case discussed in Chapter 2. I also wish to thank Dr. Osnat Herzberg for providing data for the TnC test case described in Chapter 3, and Dr. Marie Fraser for the retrieval of this data.

I am indebted to members of Dr. Michael James' laboratory for the beta testing of the likelihood target functions in X-PLOR. As well, I would like to thank Dr. Stanley Moore, Miss Katherine Bateman and Mrs. Nina Khazanovich-Bernstein for testing the likelihood functions in TNT and Dr. Paul Adams and Dr. Axel Brünger for thoroughly testing the likelihood functions in the CNS refinement package.

Finally, I wish to thank my family and in particular my mom and dad for always providing me with unconditional support.

Contents

1 Introduction **1**
- 1.1 An Overview of X-ray Crystallography . . . 1
- 1.2 Initial Phase Estimates . . . 3
- 1.3 Crystal Structure Refinement . . . 5
- 1.4 Least-Squares: A Special Case of Maximum Likelihood . . . 8

2 Applying Maximum Likelihood **9**
- 2.1 MLF1: An amplitude-based likelihood function . . . 11
- 2.2 MLF2: An intensity-based likelihood function . . . 12
- 2.3 Calibration of structure factor probabilities . . . 14
- 2.4 Test refinements . . . 15
  - 2.4.1 *Streptomyces griseus* trypsin . . . 15
  - 2.4.2 *Trypanosoma brucei* glycosomal triosephosphate isomerase . . . 17

3 Incorporating Prior Phase Information **19**
- 3.1 Deriving MLHL . . . 20
- 3.2 Test refinements . . . 24
  - 3.2.1 Troponin-C . . . 25

4 Conclusions **29**
- 4.1 Future Developments . . . 30

A Deriving P($|F|;|F_c|$) **37**

B Implementation of MLF2 **40**

C Series representation of MLHL **46**

D Phased likelihood with measurement errors **48**

List of Tables

1.1 Observation to Parameter Ratio versus Resolution 6

2.1 Refinement statistics for the SGT test case. 16

3.1 Refinement statistics for the TnC test case. 25

List of Figures

1.1 Phase ambiguity in an SIR experiment 4

2.1 R-factors through the test refinements of gTIM. 18
2.2 Phase accuracy after gTIM test refinements. 18

3.1 Plots of probability distributions for TnC test case reflection 1 1 6 . . 22
3.2 Map correlations after the TnC test refinements. 26
3.3 Combined phase SIGMAA map of the starting model. 27
3.4 Combined phase SIGMAA map of the vector residual model 27
3.5 Combined phase SIGMAA map of the MLHL model 28

List of Symbols

$\mathbf{h} = (h, k, l)$ - Miller indices

$\mathbf{x} = (x, y, z)$ - atomic positions

$|F_o|$ - experimental amplitude of the structure factor

σ_F - experimental uncertainty in amplitude of the observed structure factor

$F = |F|e^{i\alpha} = A + iB$ - true structure factor

$F_c = |F_c|e^{i\alpha_c} = A_c + iB_c$ - calculated structure factor

$J_o = |F_o|^2$

σ_j - experimental uncertainty in structure factor amplitude squared

$J = |F|^2$

$J_c = |F_c|^2$

$\alpha_{centroid}$ - centroid phase from prior experimental phase probability distribution

$A_{hl}, B_{hl}, C_{hl}, D_{hl}$ - Hendrickson-Lattman coefficients for phase probability distribution

$\mathbf{s}$ - vector of position in reciprocal space; $s = |\mathbf{s}| = 2\sin\theta/\lambda$

$\Delta\mathbf{x}$ - error in position of atoms

$D(\mathbf{s}) = < \cos(\Delta\mathbf{x} \cdot \mathbf{s}) >$ (Luzzati, 1952) ; (Read, 1990)

f_j - scattering factor for atom j

$\Sigma_N = \sum_{j=1}^{N} f_j^2$ - sum of scattering factors squared for all N atoms in a crystal

$\Sigma_P = \sum_{j=1}^{P} f_j^2$ - sum of scattering factors squared for all P atoms in a model

$\sigma_\Delta^2 = \Sigma_N - D^2\Sigma_P$

ϵ - expected intensity factor of diffracting plane (Stewart & Karle, 1976)

$I_0(x)$ and $I_1(x)$ - zero and first order modified Bessel functions of the first kind

$P(A, \ldots ; B, \ldots)$ - conditional probability distribution of $(A, \ldots)$ when $(B, \ldots)$ are known

$P(A) = \int P(A, B)dB$ - marginal probability distribution of A

Chapter 1

Introduction

Understanding biological phenomena often requires knowledge of processes at a molecular level. To fully elucidate a molecule's function in a process frequently demands its three dimensional structure. For instance, developing a drug to bind and inhibit a molecule requires precise information about the target's binding pocket. A predominant way of determining the three dimensional fold of a molecule is X-ray Crystallography. Below, a brief overview of X-ray Crystallography is given. For a more detailed discussion, please see (Blundell & Johnson, 1976) and (Drenth, 1994).

1.1 An Overview of X-ray Crystallography

X-ray Crystallography involves growing crystals of the molecule of interest, and exposing this crystal to X-ray radiation. The interaction of X-ray waves with a molecule's electrons causes diffraction of these waves. The resulting diffraction pattern produced contains information that allows for the reconstruction of the molecule's electron density.

The interaction of an X-ray wave with electrons causes scattering of the X-rays in all directions. Since a crystal is a regular repeating array of a molecule, the diffraction

of a wave by a crystal leads to both constructive and destructive interference of waves. In other words, some of the diffracted waves are constructively amplified, while some are cancelled through destructive interactions with other waves. The diffraction of waves by molecules in a crystal can be represented by an expression known as the structure factor $F(h,k,l)$.

$$F(h,k,l) = \int_0^1 \int_0^1 \int_0^1 \rho(x,y,z) \exp(2\pi i \mathbf{h} \cdot \mathbf{x}) dx dy dz \tag{1.1}$$

$\rho(x,y,z)$ represents the electron density distribution in a crystalline unit cell and $\mathbf{h} = (h,k,l)$ index the space of the diffraction pattern, also referred to as reciprocal space. The indices h,k,l are referred to as Miller indices. As mentioned above, diffraction from a crystal leads to constructive interference of certain waves. Constructive interference, and thus a spot on a diffraction pattern is observed only for integer Miller indices. Bragg has shown that the process of diffraction from a crystal can also be explained by reflection from a set of parallel planes (James, 1962). Therefore, the diffraction spots are sometimes referred to as reflections.

Although F is in general a complex number, if the density ρ is centrosymmetric the imaginary part vanishes and F reduces to a real number. However, even for crystallized asymmetric objects, there can exist a class of reflections that gives information only about a centrosymmetric projection of a crystal. In this class of reflections, known as the centric reflections, the phase choice of F is restricted to two possible values separated by 180 degrees. A reflection that is not centric is referred to as an acentric reflection. Acentric reflections make up the bulk of the data collected from a protein crystal.

Mathematically, a crystal is modelled by dividing three dimensional space into a regular repeating array of cells. Each "unit cell" contains some, all, or multiple copies of the molecule of interest. The molecule is characterized by its electron density

ρ. Since no substance has been found to focus X-ray beams, a molecule's electron density can not be directly observed from the X-ray diffraction experiment. However, the density ρ can be recovered from the structure factors via the inverse Fourier transform.

$$\rho(x,y,z) = \frac{1}{V}\sum_h\sum_k\sum_l F(h,k,l)\exp(-2\pi i\mathbf{h}\cdot\mathbf{x}) \tag{1.2}$$

As mentioned above, each structure factor F is a complex number. However, from the diffraction experiment, only the modulus $|F|$ can be estimated and no direct measurement of the phase is possible. Since both phases and amplitudes are needed, Fourier inversion is not possible. This hindrance is known as the phase problem of crystallography. Furthermore, the measurements of the amplitudes $|F|$ are subject to random error.

1.2 Initial Phase Estimates

To build an initial model for a crystal structure, estimates for the phases are needed. Two ways for obtaining initial estimates of phases are Molecular Replacement (MR) and Multiple Isomorphous Replacement (MIR).

Often, protein structures exhibit similar folds to other proteins that can be predicted by the sequence similarity of the proteins. The Molecular Replacement method exploits this property of proteins by obtaining initial phases for a molecule of unknown structure from a related molecule of known structure.

The method of Multiple Isomorphous Replacement involves adding a "heavy atom" or an atom with a high atomic number to the crystal in order to perturb the diffraction pattern of the crystal. If the addition of the heavy atom does not disrupt the protein structure or the crystal packing, then phase information can be

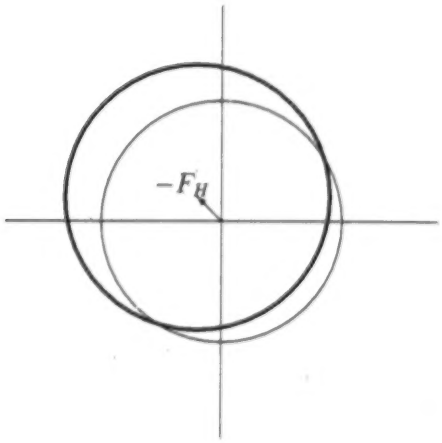


Figure 1.1: Phase ambiguity in an SIR experiment
The thin curve represents a circle centered at the origin of radius $|F_P|$ and the thick curve is a circle of radius $|F_{PH}|$ centered at $-F_H$.

obtained. The crystal containing the protein structure and the heavy atom is called a derivative. To determine the phase of a structure factor, the original or native ($|F_P|$) and the derivative ($|F_{PH}|$) structure factor amplitudes must be measured. Furthermore, the coordinates of the heavy atom and its corresponding structure factor (F_H) must be determined. The native structure factor amplitude restricts the structure factor, F_P, to a circle in the complex plane. As well, an independent source for F_P can be obtained from a circle centered at the end of the vector $-F_H$ of radius $|F_{PH}|$, since $F_P = F_{PH} - F_H$.

Figure 1.1 gives an example of phasing from a Single Isomorphous Replacement (SIR) experiment. In this case, a circle of radius $|F_P|$ centered at the origin intersects twice with a circle of radius $|F_{PH}|$ centered at $-F_H$. The two intersection points of these curves correspond to the two possible phase values for F_P from this experiment. This phase ambiguity can be resolved by measuring the amplitude of a different heavy

atom derivative.

The inclusion of these initial phase estimates with the observed structure factor amplitudes allows for the construction of an initial model. Because of the errors associated with measuring structure factor amplitudes and the errors associated with phasing a model, the initial model constructed often contains errors in its atomic coordinates, missing atoms, or extra atoms. Therefore, to obtain the most accurate model, refining the initial model against all of the available experimental data is essential.

1.3 Crystal Structure Refinement

To obtain the most accurate possible crystal structure, one typically refines the atomic model to optimize its agreement with the observed diffraction data. The standard macromolecular refinement programs, PROLSQ (Konnert & Hendrickson, 1980), TNT (Tronrud, Ten Eyck & Matthews, 1987), X-PLOR (Brünger, Kuriyan & Karplus, 1987), and GROMOS (Fujinaga, Gros & van Gunsteren, 1989), minimize a residual that is the weighted sum of squared deviations between the observed ($|F_o|$) structure factor amplitude and the structure factor amplitude calculated ($|F_c|$) from an atomic model of a protein using equation (1.1), including a relative scale factor k and a weighting factor w_h:

$$\sum_h w_h(|F_o| - k|F_c|)^2 \tag{1.3}$$

The refinement programs differ primarily in their minimization methods. Conjugate gradients is a common method for finding the local minima of a function and employs an iterative cyclical procedure. The first cycle involves determining an optimal search or step direction and the next cycles attempt to find the best step length in this

Table 1.1: Observation to Parameter Ratio versus Resolution

Resolution (Å)	Observations/parameter
3.5	0.5
3.0	0.8
2.5	1.4
2.0	2.8
1.5	6.2

direction. This process is repeated until the function's gradient nears the zero vector.

The Cartesian coordinates (x, y, z) and a thermal motion parameter or B-value are typically used to parameterize an atom for a given model. Thus, the calculated structure factor F_c is a function of x, y, z and B for all the atoms in a model.

The quality and amount of diffraction data collected depends largely on the quality of crystals obtained. High angle diffraction spots or high resolution data can be seen for good crystals, whereas poorer quality crystals only diffract to low resolution. Table 1.1 shows the observation to parameter ratio as a function of resolution for a protein crystal with a typical packing density and assuming four parameters x, y, z and B for an atom.

Since most macromolecular refinements have an unfavourable parameter to observation ratio, geometrical restaints representing prior information of ideal molecule geometry (ie. ideal bond angles and bond lengths) are added. Alternatively, one can reduce the number of parameters using constrained models to improve the unfavourable ratio.

Naturally, the quality of the model after refinement depends on the X-ray target being used. Experience has shown that the above least-squares target works poorly for very incomplete models and models with large coordinate errors. Furthermore, the least-squares function does not account for the effect of the observed structure

factor amplitude's measurement error and can not justifiably be generalized to include prior phase information. In order to overcome these shortcomings, a more general maximum likelihood approach should be considered, as suggested by Read (1990) and Bricogne (1991; 1993).

The principle of maximum likelihood formalizes the idea that the quality of a model is judged by its consistency with the observations. To say that a model is consistent with an observation means that, if the model were correct, there would be a reasonably high probability of making an observation with that value. Taking the relevant observations as a set, then, the probability of making the entire set of observations is an excellent measure of the quality of the model. If we assume that the observations are independent, the joint probability of making the set of observations is the product of the probabilities of making each independent observation. This joint probability is the likelihood function (L):

$$L = \prod_{hkl} P(|F_o|; |F_c|, \alpha_c). \tag{1.4}$$

where $P(|F_o|; |F_c|, \alpha_c)$ is the conditional probability of the observation $|F_o|$ given the calculated amplitude $|F_c|$ and phase α_c. Since it is more convenient to work with sums than products, one typically works with the logarithm of the likelihood function. As well, the maximization problem can be turned into a minimization problem by multiplying by negative one. Therefore, defining $\mathcal{L} = -\log(L)$ gives the following:

$$\mathcal{L} = -\sum_{hkl} \log(P(|F_o|; |F_c|, \alpha_c)) \tag{1.5}$$

1.4 Least-Squares: A Special Case of Maximum Likelihood

The least-squares refinement target could be considered to arise from the principle of maximum likelihood, if the expected value of $|F_o|$ were $k|F_c|$, and the probability distribution of $|F_o|$ given $|F_c|$ were a Gaussian, and the standard deviation were independent of the parameters of the atomic model.

As will be shown, none of the above assumptions are true. For this reason, we should return to first principles and apply a maximum likelihood analysis to the problem of crystal structure refinement.

In the case of crystallographic refinement, it is not strictly true that the diffraction observations are independent; if they were, direct methods and density modification would not work. There is doubtless much useful information to be gained by working with higher order collections of structure factors (Bricogne, 1993), but useful results are obtained even when independence is assumed, as will be shown.

Chapter 2

Applying Maximum Likelihood

To apply maximum likelihood, one must start from the probability of making a measurement, given the model, its errors, and the measurement errors. It has been shown previously that various sources of random error in the model have equivalent effects on the probability distribution for the true structure factor, whether the errors are in atomic positions or temperature factors or whether there are missing or extra atoms: in each case the distribution of the true structure factor is well approximated by a Gaussian distribution centered on DF_c (Read, 1990). The parameter D, a function of the reciprocal space vector $\mathbf{s}$, is the Fourier transform of the probability distribution of the coordinate error (Δx) (Luzzati, 1952); (Read, 1990) and intuitively represents the fraction of the calculated structure factor that is correct. In the case of acentric structure factors, which make up the bulk of data for macromolecular structures, the distribution ($P_a(F;F_c)$) is a two-dimensional Gaussian in the complex plane, while for centric structure factors, it is a one-dimensional Gaussian ($P_c(F;F_c)$):

$$P_a(F;F_c) = \frac{1}{\pi\epsilon\sigma_\Delta^2} e^{-\frac{|F-DF_c|^2}{\epsilon\sigma_\Delta^2}} \tag{2.1}$$

[1] A version of this chapter has been published. Pannu and Read (1996) *Acta Cryst* **A52: 659-668.**

$$P_c(F;F_c) = \frac{1}{\sqrt{2\pi\epsilon\sigma_\Delta^2}} e^{-\frac{|F-DF_c|^2}{2\epsilon\sigma_\Delta^2}} \tag{2.2}$$

In Appendix A, the probability of the true structure factor amplitude ($|F|$), conditional on the calculated amplitude ($|F_c|$), is shown to be the following for the acentric and centric case, respectively:

$$P_a(|F|;|F_c|) = \frac{2|F|}{\epsilon\sigma_\Delta^2} e^{-\frac{|F|^2+D^2|F_c|^2}{\epsilon\sigma_\Delta^2}} \mathrm{I}_0\left(\frac{2|F|D|F_c|}{\epsilon\sigma_\Delta^2}\right) \tag{2.3}$$

$$P_c(|F|;|F_c|) = \sqrt{\frac{2}{\pi\epsilon\sigma_\Delta^2}} e^{-\frac{|F|^2+D^2|F_c|^2}{2\epsilon\sigma_\Delta^2}} \cosh\left(\frac{|F|D|F_c|}{\epsilon\sigma_\Delta^2}\right) \tag{2.4}$$

The probability distribution required to apply maximum likelihood, however, is the probability of the observed diffraction measurement given the calculated diffraction measurement, as the true value is not known. We have used two methods to approximate this distribution, differing in the level of approximation and in the distribution assumed for the observational error. In the first method (MLF1), the measurement error is assumed to be Gaussian in structure factor amplitudes, and a Gaussian approximation is made for the resultant combined distribution, expressed in terms of structure factor amplitudes. In the second method (MLF2), the measurement error is assumed to be Gaussian in the intensities, and a series representation of the resultant combined distribution is expressed in terms of structure factor amplitudes squared.

2.1 MLF1: An amplitude-based likelihood function

If the probability of the measurement error ($P(|F_o| - |F|)$) is assumed to be Gaussian in structure factor amplitudes with standard deviation σ_F , then the required probability distribution $P(|F_o|; |F_c|)$ is obtained by convoluting $P(|F|; |F_c|)$ by $P(|F_o| - |F|)$.

$$P(|F_o|; |F_c|) = P(|F|; |F_c|) \otimes P(|F_o| - |F|) \tag{2.5}$$

As far as we have been able to determine, there is no exact analytical solution to this convolution for the important acentric case. However, a good Gaussian approximation can be obtained using the first two central moments of the distribution. The expected value for the acentric case is given by the following:

$$<|F_o|>= \frac{\sqrt{\pi\epsilon\sigma_\Delta^2}}{2}\ \Phi(-\frac{1}{2}, 1, -\frac{D^2|F_c|^2}{\epsilon\sigma_\Delta^2}) \tag{2.6}$$

For the centric case, the expected value is

$$<|F_o|>= \sqrt{\frac{2\epsilon\sigma_\Delta^2}{\pi}}\ \Phi(-\frac{1}{2}, \frac{1}{2}, -\frac{D^2|F_c|^2}{2\epsilon\sigma_\Delta^2}) \tag{2.7}$$

In these expressions, $\Phi(a, b, x)$ is Kummer's Confluent Hypergeometric Function (Luke, 1977), also denoted by ${}_1\mathrm{F}_1(a, b, x)$. The variance for both the acentric and centric distributions is given by the following:

$$\sigma_{ML}^2 = \epsilon\sigma_\Delta^2 + \sigma_F^2 + D^2|F_c|^2 - <|F_o|>^2 \tag{2.8}$$

As $|F_c|$ increases, σ_{ML}^2 tends towards $\epsilon\sigma_\Delta^2 + \sigma_F^2$ in the centric case, or $\frac{1}{2}\epsilon\sigma_\Delta^2 + \sigma_F^2$ in the acentric case because, in the limit, only the component of model error parallel

to $\mathbf{F_c}$ contributes to the error in the amplitude. When these moments are used to construct a Gaussian approximation, the negative log likelihood function ($\mathcal{L}$) is

$$\mathcal{L} = \sum_{hkl} \frac{1}{2}\log(2\pi) + \log(\sigma_{ML}) + \frac{1}{2\sigma_{ML}^2}(|F_o| - <|F_o|>)^2 \tag{2.9}$$

Eliminating the constant term $\frac{1}{2}\log(2\pi)$ gives the function minimized in refinement.

$$\sum_{hkl} \log(\sigma_{ML}) + \frac{1}{2\sigma_{ML}^2}(|F_o| - <|F_o|>)^2 \tag{2.10}$$

2.2 MLF2: An intensity-based likelihood function.

The second method that we use to derive the required probability distribution works in terms of structure factor amplitudes squared ($J = |F|^2$). Two advantages are attained by working in J instead of F. First, measurement errors frequently lead to a negative net intensity, which is reduced to negative J: when these legitimate observations are transformed to $|F|$, one has the choice of omitting them, replacing them with zero, or replacing them with a non-zero Bayesian posterior value (French & Wilson, 1978) . By working in terms of J, this problem is avoided. Furthermore, a Gaussian measurement error is better justified in J, than in $|F|$. In principle, maximum likelihood is insensitive to variable transformations such as from $|F|$ to $|F|^2$ (Edwards, 1992). If MLF2 did not differ from MLF1 in the distribution assumed for the measurement error, the two likelihood functions would differ only in the precision of the approximation.

The required probability distribution $P(J_o; J_c)$ is derived by multiplying $P(J; J_c)$ with the Gaussian probability of the measurement error ($P(J_o; J)$) with standard

deviation σ_j, and integrating over the true structure factor amplitude squared (J).

$$P(J_o; J_c) = \int_0^\infty P(J_o; J) \times P(J; J_c) dJ \tag{2.11}$$

A series representation of $P(J_o; J_c)$ can be computed. For acentric reflections the distribution is the following:

$$P_a(J_o; J_c) = \frac{1}{\sqrt{2\pi\epsilon\sigma_\Delta^2}} e^{-\frac{J_o^2}{2\sigma_j^2} - \frac{D^2 J_c}{\epsilon\sigma_\Delta^2}} \sum_{n=0}^{\infty} \left(\frac{D^2 J_c \sigma_j}{\epsilon^2 \sigma_\Delta^4}\right)^n \frac{1}{n!} e^{\frac{(\sigma_j^2 - J_o\epsilon\sigma_\Delta^2)^2}{4\epsilon^2\sigma_\Delta^4\sigma_j^2}} \mathrm{D}_{-n-1}\left(\frac{\sigma_j^2 - J_o\epsilon\sigma_\Delta^2}{\epsilon\sigma_\Delta^2\sigma_j}\right) \tag{2.12}$$

$\mathrm{D}_{-n-1}(x)$ is a parabolic cylinder function. For centric reflections, the distribution is given below.

$$P_c(J_o; J_c) = \frac{1}{2\sqrt{\pi\sigma_j}\epsilon\sigma_\Delta} e^{-\frac{J_o^2}{2\sigma_j^2} - \frac{D^2 J_c}{2\epsilon\sigma_\Delta^2}} \sum_{n=0}^{\infty} \left(\frac{D^2 J_c \sigma_j}{2\epsilon^2 \sigma_\Delta^4}\right)^n \frac{1}{(2n)!!} e^{\frac{(\sigma_j^2 - 2J_o\epsilon\sigma_\Delta^2)^2}{16\epsilon^2\sigma_\Delta^4\sigma_j^2}} \mathrm{D}_{-n-\frac{1}{2}}\left(\frac{\sigma_j^2 - 2J_o\epsilon\sigma_\Delta^2}{2\epsilon\sigma_\Delta^2\sigma_j}\right) \tag{2.13}$$

After eliminating terms that are constant within a cycle of refinement, the negative log likelihood ($\mathcal{L}$) for the acentric case is the following:

$$\mathcal{L} = \sum_{hkl} \log(\epsilon\sigma_\Delta^2) + \frac{D^2 J_c}{\epsilon\sigma_\Delta^2} - \log\left(\sum_{n=0}^{\infty} \left(\frac{D^2 J_c \sigma_j}{\epsilon^2 \sigma_\Delta^4}\right)^n \frac{1}{n!} e^{\frac{(\sigma_j^2 - J_o\epsilon\sigma_\Delta^2)^2}{4\epsilon^2\sigma_\Delta^4\sigma_j^2}} \mathrm{D}_{-n-1}\left(\frac{\sigma_j^2 - J_o\epsilon\sigma_\Delta^2}{\epsilon\sigma_\Delta^2\sigma_j}\right)\right) \tag{2.14}$$

and for centric reflections the negative log likelihood expression is given below.

$$\mathcal{L} = \sum_{hkl} \frac{1}{2} \log(\epsilon\sigma_\Delta^2) + \frac{D^2 J_c}{2\epsilon\sigma_\Delta^2} - \log\left(\sum_{n=0}^{\infty} \left(\frac{D^2 J_c \sigma_j}{2\epsilon^2 \sigma_\Delta^4}\right)^n \frac{1}{(2n)!!} e^{\frac{(\sigma_j^2 - 2J_o\epsilon\sigma_\Delta^2)^2}{16\epsilon^2\sigma_\Delta^4\sigma_j^2}} \mathrm{D}_{-n-\frac{1}{2}}\left(\frac{\sigma_j^2 - 2J_o\epsilon\sigma_\Delta^2}{2\epsilon\sigma_\Delta^2\sigma_j}\right)\right) \tag{2.15}$$

Equations (2.14)-(2.17) are derived in Appendix B.

To optimize refinement targets, derivatives with respect to F_c are commonly used. Calculating the derivative of a least-squares target on amplitudes results in a division by $|F_c|$ that can result in a singularity (Schwarzenbach et al., 1989). Both MLF1 and MLF2 functions eliminate this singularity.

2.3 Calibration of structure factor probabilities

The value of the likelihood function depends on the parameters of the atomic model. It also depends on the resolution-dependent parameters D and σ_Δ^2, which characterize the effect of model error on the structure factor probability distributions. (In fact D and σ_Δ^2 are not independent and can each be computed from the single parameter σ_A (Read, 1990).) In principle, it would be best to optimize the likelihood function by adjusting all parameters simultaneously, including coordinates, B-factors *and* σ_A values. Unfortunately, a problem arises if the σ_A values are refined using the same data against which the model is refined: the poor parameter to observation ratio allows overfitting of the amplitudes, which results in an overestimation of σ_A and hence an underestimation of the errors in the calculated structure factors (Lunin & Urzhumtsev, 1984; Read, 1986). This leads to a positive feedback cycle in which the pressure to overfit becomes stronger. In our first attempt to implement maximum likelihood refinement, this problem was ignored. As the quality of the likelihood function depends strongly on the accuracy of σ_A estimates, the results were unimpressive.

The solution adopted is to use cross-validation data (a minority of data omitted from the refinement target) in an active way to provide unbiased estimates of structure factor accuracy. These data are normally used to compute R_{free}, an unbiased measure of refinement progress (Brünger, 1992). The use of cross-validation data to estimate

σ_A is complicated, however, by the fact that stable estimates require 500 to 1000 reflections in each resolution shell, especially when the true value is low (Read, 1986). To overcome the problem of instability, we exploit the fact that σ_A varies smoothly with resolution. A simple correction, in which a penalty is applied when a σ_A value lies far from the line connecting its two neighbours, is sufficient (Read, unpublished).

A better solution would be to refine the σ_A values as parameters in the refinement, but to make allowance for the fact that they are biased estimates, in using them in the likelihood function. Lacking a theoretical basis for the correction for bias, however, this solution cannot yet be applied. We are currently studying the effect of refinement bias on the structure factor distributions, to lay the groundwork for such an improved treatment.

2.4 Test refinements

The two maximum likelihood targets have been implemented in the programs CNS, TNT (Tronrud, Ten Eyck & Matthews, 1987) and X-PLOR (Brünger, Kuriyan & Karplus, 1987). Results from runs of the modified X-PLOR on two test systems will be discussed here. In each test, the suggested weighting factor (WA) for the diffraction terms in the target, obtained by comparing the gradients from the diffraction and energy terms (Brünger, Karplus & Petsko, 1989), was divided by two.

2.4.1 *Streptomyces griseus* trypsin

The crystal structure of *Streptomyces griseus* trypsin (Read & James, 1988) (SGT) was solved originally by molecular replacement, using the structure of bovine trypsin (Chambers & Stroud, 1979) (BT) as a search model. In order to compare the power of the maximum likelihood and least-squares targets in a case where the phase errors

Table 2.1: Refinement statistics for the SGT test case.

	Start	Least-squares	MLF1	MLF2
R-factor	0.515	0.403	0.416	0.422
R_{free}	0.542	0.511	0.525	0.528
Mean phase error	62.2	60.0	56.7	56.5
Mean cos(phase error)	0.365	0.394	0.436	0.437

are known exactly, we used data calculated from SGT as error-free amplitudes $|F_o|$, and a superimposed model of BT as a starting structure. Since these two proteins share about 33 % sequence identity, BT provides a relatively poor model that will only be capable of refining into a local minimum.

Data from infinity to 2.8 Å resolution (5732 reflections, of which 578 were flagged as cross-validation data) were used for both refinements. (One often omits the low resolution data for least-squares refinement because of the complications caused by disordered solvent, but in this case there is no disordered solvent). In total, 420 cycles of energy minimization refinement in X-PLOR were carried out. Table 2.1 shows the results obtained in the different refinements.

While none of the refinements could achieve an accurate model, owing to the inadequacies of the starting model, the maximum likelihood targets gave more than twice as large an improvement in the average phase error. Note that, owing probably to the small number of reflections used in this case, R_{free} provides a weak indication of phase accuracy.

2.4.2 *Trypanosoma brucei* glycosomal triosephosphate isomerase

At an intermediate stage in the refinement of the glycosomal triosephosphate isomerase (gTIM) from *Trypanosoma brucei* (Wierenga, Noble, Vriend, Nauche & Hol, 1991), data to a resolution of 1.83 Å became available to replace the data to 2.4 Å resolution that had been used to that point (Wierenga, Kalk & Hol, 1987). We tested the three refinement targets on this intermediate model, using the observed diffraction data (model and data kindly supplied by Dr. R.K. Wierenga). Of 38812 observed amplitudes, 1014 were flagged randomly as cross-validation data. Because this is a real data set measured from a crystal with disordered solvent, data from infinity to 8.0 Å resolution were omitted in the least-squares refinement, while they were used in both maximum likelihood refinements. In each case, 250 cycles of energy minimisation (EM) refinement were run, followed by 30 cycles of B-factor refinement.

As shown in Figures 2.1 and 2.2, both maximum likelihood target functions achieved a significantly greater improvement in the model, measured by both R_{free} and phase differences with the final model.

As one might expect from the increased precision of the approximation, the MLF2 target gives significantly better results than MLF1. This improvement is achieved for a modest computational cost. Compared to an equivalent refinement with the least-squares target, the MLF1 target requires about 1% more computer time, while the MLF2 target requires about 10% more computer time.

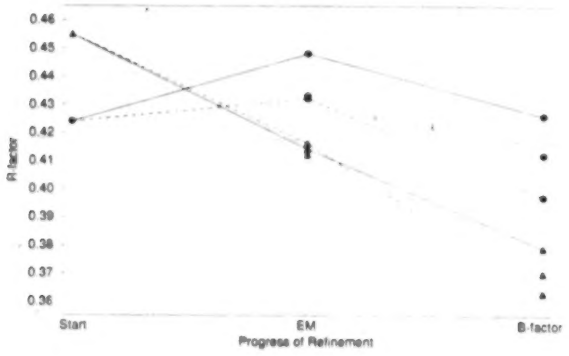


Figure 2.1: R-factors through the test refinements of gTIM.
The solid lines indicate R-factors for the least-squares target, the dotted lines indicate R-factors for the MLF1 target, and the dashed lines indicate R-factors for the MLF2 target. R_{free} values for the three different target functions are represented by circles, and R values for the three different target functions are represented by triangles.

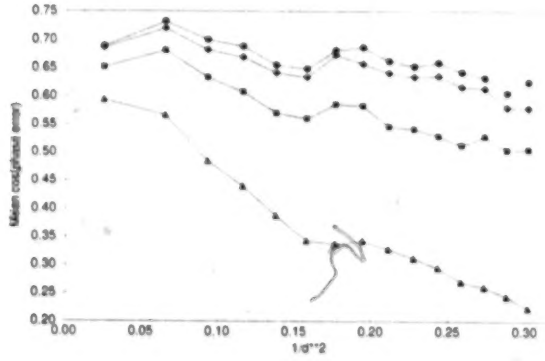


Figure 2.2: Phase accuracy after gTIM test refinements.
The phase accuracy is computed as the mean cosine of the phase error, which is comparable to the mean figure of merit. Triangles correspond to the starting model, squares to the least-squares model, diamonds to the MLF1 model, and circles to the MLF2 model.

Chapter 3

Incorporating Prior Phase Information

In order to improve the parameter to observation ratio, inclusion of prior phase information in the refinement protocol has been previously proposed. A number of methods have been derived to incorporate this additional source of information. One method involves adding a square well potential around the centroid phase to the least-squares target. The vector residual represents another target function used to include prior phase information.

$$V_{residual} = \sum_{hkl}(|F_o|\cos(\alpha_{centroid}) - A_{calc})^2 + (|F_o|\sin(\alpha_{centroid}) - B_{calc})^2 \qquad (3.1)$$

where A_{calc} and B_{calc} represent the real and imaginary components of the calculated structure factor and $\alpha_{centroid}$ represents the expected value of the phase. Typically, prior phase information for macromolecules is not very accurate, and since the above two target functions do not consider the errors in the phase measurements, both are not theoretically justified.

Although the previous chapter details initial results that are striking, the maxi-

mum likelihood method allows for a rational incorporation of other sources of information which should lead to further improvements. A likelihood function has been derived that incorporates experimental phase information frequently determined in a crystal structure determination. A likelihood function incorporating prior phase information has also been proposed by Bricogne and Irwin (1996) and Murshudov, Dodson, and Vagin (1996). The function, MLHL, or Maximum Likelihood function using Hendrickson-Lattman coefficients (Hendrickson & Lattman, 1970) has been implemented in the refinement programs CNS, TNT (Tronrud, Ten Eyck & Matthews, 1987), and X-PLOR (Brünger, Kuriyan & Karplus, 1987).

3.1 Deriving MLHL

Intuitively, the derivation of the MLHL target function follows similarly from the derivation of a maximum likelihood function lacking prior phase information. In either case $P(|F_o|; |F_c|, \alpha_c)$ is derived from the joint probability of the true structure factor and the calculated structure factor, denoted $P(F, F_c)$. However, in the case of a likelihood function lacking prior phase information, $P(|F_o|; |F_c|, \alpha_c)$ is obtained by integrating $P(F, F_c)$ uniformly over all possible phases. In the derivation of MLHL, the distribution $P(|F_o|; |F_c|, \alpha_c)$ is determined by an integration over all possible phases of $P(F, F_c)$ weighted by an experimental prior phase probability distribution.

In order to derive the MLHL function mathematically, the distribution $P(|F|, \Delta\alpha; |F_c|, \alpha_c)$ is needed. Appendix A gives this distribution for acentric and centric reflections, respectively.

$$P_a(|F|, \Delta\alpha; |F_c|, \alpha_c) = \frac{|F|}{\pi\epsilon\sigma_\Delta^2} \exp\left(\frac{-|F|^2 - D^2|F_c|^2 + 2|F|D|F_c|\cos(\Delta\alpha)}{\epsilon\sigma_\Delta^2}\right) \qquad (3.2)$$

$$P_c(|F|, \Delta\alpha; |F_c|, \alpha_c) = \frac{1}{\sqrt{2\pi\epsilon\sigma_\Delta^2}} \exp\left(\frac{-|F|^2 - D^2|F_c|^2 + 2|F|D|F_c|\cos(\Delta\alpha)}{2\epsilon\sigma_\Delta^2}\right) \quad (3.3)$$

In the above equations, $\Delta\alpha$ is the phase difference between the true phase and the calculated phase. Hendrickson and Lattman (1970) have shown that the prior probability distribution of a phase (α) can be represented in the following form:

$$P(\alpha) = N \exp\{A_{hl}\cos(\alpha) + B_{hl}\sin(\alpha) + C_{hl}\cos(2\alpha) + D_{hl}\sin(2\alpha)\} \quad (3.4)$$

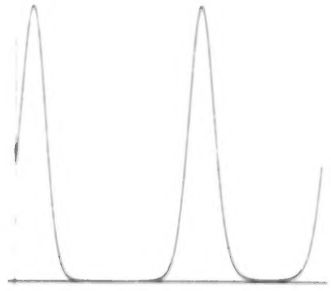
where A_{hl}, B_{hl}, C_{hl} and D_{hl} are Hendrickson-Lattman coefficients and N is a normalization constant.

In the acentric case, multiplication of the distribution $P_a(|F|, \Delta\alpha; |F_c|, \alpha_c)$ with the prior probability distribution $P(\alpha)$ gives the joint probability distribution $P_a(|F|, \Delta\alpha, \alpha; |F_c|, \alpha_c)$. Integrating the true phase out of this joint probability distribution gives the required distribution.

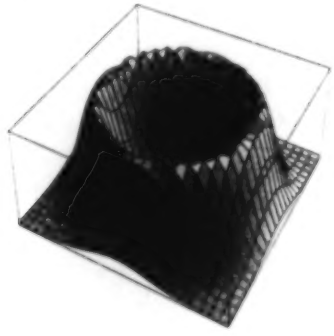
$$\begin{aligned}
P(|F|; |F_c|, \alpha_c) &= \int_0^{2\pi} P_a(|F|, \Delta\alpha; |F_c|, \alpha_c) P(\alpha) d\alpha \\
&= \int_0^{2\pi} P_a(|F|, \alpha - \alpha_c; |F_c|, \alpha_c) P(\alpha) d\alpha \\
&= \frac{N|F|}{\pi\epsilon\sigma_\Delta^2} \exp\left(\frac{-|F|^2 - D^2|F_c|^2}{\epsilon\sigma_\Delta^2}\right) \\
&\quad \int_0^{2\pi} \exp\{(A_{hl} + \frac{2|F|D|F_c|\cos(\alpha_c)}{\epsilon\sigma_\Delta^2})\cos(\alpha) + \\
&\quad (B_{hl} + \frac{2|F|D|F_c|\sin(\alpha_c)}{\epsilon\sigma_\Delta^2})\sin(\alpha) + C_{hl}\cos(2\alpha) + D_{hl}\sin(2\alpha)\} d\alpha
\end{aligned} \quad (3.5)$$

A surface plot of equation 3.5 is shown in Figure 3.1. Taking the minus logarithm of equation 3.5, removing all terms that are constant, and summing over all reflections

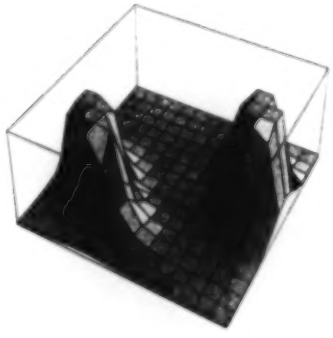
Figure 3.1: Plots of probability distributions for TnC test case reflection 1 1 6



3.1(a) Phase probability curve indicating the two most probable phase choices.



3.1(b) Surface plot of $P(|F_o|;|F_c|,\alpha_c)$ lacking prior phase information versus the real and imaginary part of F_c. This distribution is radially symmetric, thus the extremum chosen will be in the direction of the model phase.



3.1(c) Surface plot of $P(|F_o|;|F_c|,\alpha_c)$ incorporating prior phase information versus the real and imaginary part of F_c. This distribution reinforces both phase choices seen in the SIR experiment as peaks in the function.

gives the desired target function.

$$\mathcal{L} = \sum_{hkl} \frac{|F|^2 + D^2|F_c|^2}{\epsilon\sigma_\Delta^2} - \log\Big(\int_0^{2\pi} \exp\{(A_{hl} + \frac{2|F|D|F_c|\cos(\alpha_c)}{\epsilon\sigma_\Delta^2})\cos(\alpha) + (B_{hl} + \frac{2|F|D|F_c|\sin(\alpha_c)}{\epsilon\sigma_\Delta^2})\sin(\alpha) + C_{hl}\cos(2\alpha) + D_{hl}\sin(2\alpha)\} d\alpha \Big) \tag{3.6}$$

A series representation for the integral in equation 3.6 exists and is given in Appendix C. Unfortunately this series exhibits numerical instabilities for particular arguments, so the above integral is evaluated numerically in the general case of non-zero A_{hl}, B_{hl}, C_{hl}, and D_{hl} Hendrickson-Lattman coefficients. However, in the special case when C_{hl} and D_{hl} are both zero, an analytical form exists:

$$\mathcal{L} = \sum_{hkl} \frac{|F|^2 + D^2|F_c|^2}{\epsilon\sigma_\Delta^2} - \log\left\{ \mathrm{I}_0\Big(\sqrt{(\mathrm{A_{hl}} + \frac{2|\mathrm{F}|\mathrm{D}|\mathrm{F_c}|\cos(\alpha_c)}{\epsilon\sigma_\Delta^2})^2 + (\mathrm{B_{hl}} + \frac{2|\mathrm{F}|\mathrm{D}|\mathrm{F_c}|\sin(\alpha_c)}{\epsilon\sigma_\Delta^2})^2}\Big) \right\} \tag{3.7}$$

Equation number 3.7 demonstrates an important property of the MLHL function. In the case of no phase information, when all the Hendrickson-Lattman coefficients are zero, the MLHL target reduces to the minus logarithm of the Rice distribution: a maximum likelihood target function lacking prior phase information shown in Appendix A and equation 2.3.

For centric reflections, the required distribution is obtained by multiplication of the density $P_c(|F|, \Delta\alpha; |F_c|, \alpha_c)$ with the prior probability distribution $P(\alpha)$ and summing

over the two possible phase values.

$$\begin{aligned} P(|F|;|F_c|,\alpha_c) &= \sum P(|F|,\Delta\alpha;|F_c|,\alpha_c)P(\alpha) \\ &= \sum P(|F|,\alpha-\alpha_c;|F_c|,\alpha_c)P(\alpha) \\ &= \sqrt{\frac{2}{\pi\epsilon\sigma_\Delta^2}}\exp\left\{\frac{-|F|^2-D^2|F_c|^2}{2\epsilon\sigma_\Delta^2}\right\}\times \\ &\quad \cosh\left\{A_{hl}\cos(\alpha_c)+B_{hl}\sin(\alpha_c)+\frac{|F|D|F_c|}{\epsilon\sigma_\Delta^2}\right\} \end{aligned} \tag{3.8}$$

The sums in equation 3.8 are over the two values of α : α_c and $\alpha_c+\pi$. The minus log of equation 3.8 is the following:

$$\begin{aligned} \mathcal{L} &= \sum_{hkl}\frac{|F|^2+D^2|F_c|^2}{\epsilon\sigma_\Delta} \\ &\quad \log(\cosh\left\{A_{hl}\cos(\alpha_c)+B_{hl}\sin(\alpha_c)+\frac{|F|D|F_c|}{\epsilon\sigma_\Delta^2}\right\}) \end{aligned} \tag{3.9}$$

The above derivation of the MLHL function neglects the effect of measurement errors on the structure factor probability distribution. Appendix D considers the effect of measurement error in deriving the MLHL function.

3.2 Test refinements

The maximum likelihood target MLHL has been implemented in the programs CNS, TNT (Tronrud, Ten Eyck & Matthews, 1987) and X-PLOR (Brünger, Kuriyan & Karplus, 1987). Results from runs of the modified X-PLOR on one test system will be discussed here. In each test, the suggested weighting factor (WA) for the diffraction terms in the target, obtained by comparing the gradients from the diffraction and energy terms (Brünger, Karplus & Petsko, 1989), was divided by two.

Table 3.1: Refinement statistics for the TnC test case.

	Start	Least-squares	MLF1	MLF2	Vector	MLHL
R-factor	0.571	0.416	0.428	0.403	0.468	0.359
R_{free}	0.559	0.532	0.490	0.482	0.481	0.399
Mean phase error	73.7	66.0	52.0	49.7	49.5	33.4
Mean cos(phase error)	0.21	0.31	0.49	0.52	0.52	0.72
Mean map correlation	0.369	0.502	0.665	0.704	0.692	0.860

3.2.1 Troponin-C

In this test, we refined a "scrambled" starting model using only poor SIR phases to supplement the likelihood function. The test protein was troponin-C (TnC) which was originally solved at 2.8 Å resolution using MIR phases from eleven derivatives (Herzberg & James, 1985). (MIR data kindly supplied by Osnat Herzberg, with assistance from Marie Fraser.) Of these eleven derivatives, a single derivative ($TmCl_3$) was chosen. The originally determined heavy atom parameters for this derivative were further refined by MLPHARE (Otwinowski, 1991) which subsequently generated the Hendrickson-Lattman coefficients used by MLHL, and the "best" phase and figure of merit used by the vector residual. $TmCl_3$ phases were relatively poor, as MLPHARE reported a mean figure of merit of 0.39, while the mean cosine of the phase difference with the phases computed from the final published structure was 0.29.

A starting model was generated by "scrambling" (Rice & Brünger, 1994) or performing a molecular dynamics run using a target function without reference to X-ray information. The starting model generated had a root mean squared deviation of 2.28 Å with the published structure. Of the 3868 observed native reflections, 496 were flagged as cross-validation data for σ_A estimation (Read, 1997) and R-free calculation (Brünger, 1992). The test refinement involved 420 cycles of conjugate gradient refinement in X-PLOR using MLHL, the vector residual, MLF2, MLF1, and

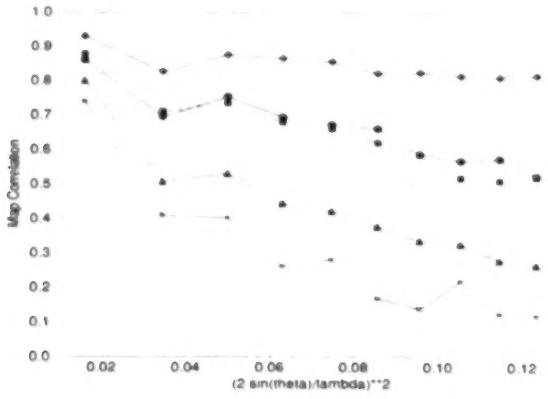


Figure 3.2: Map correlations after the TnC test refinements.
Stars correspond to the starting model, triangles to the least-squares model, circles to the MLF2 model, squares to the Vector-Residual model, and diamonds to the MLHL model.

least-squares.

Results from this test are shown in Table 3.1 and Figure 3.2. As indicated by the map correlation with the final model, MLHL clearly performed better than any other target function. As well, MLHL gave the lowest R-free value.

Figures 3.3, 3.4 and 3.5 show combined phase SIGMAA electron density maps (Read, 1997) for a region of TnC of the starting, vector residual, and MLHL models and SIR phases used in refinement.

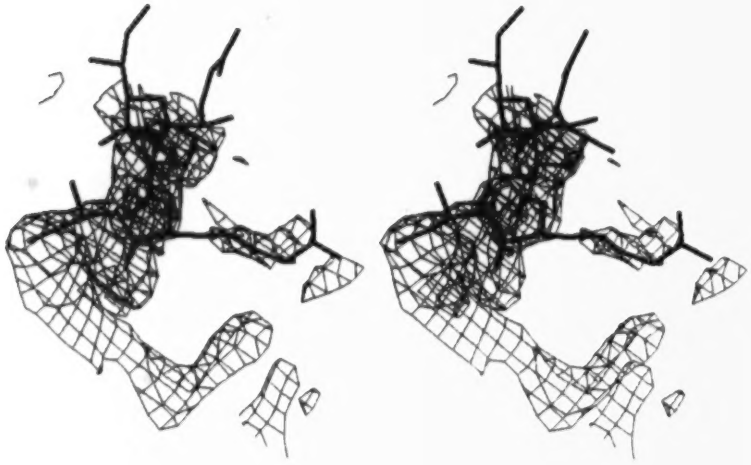


Figure 3.3: Combined phase SIGMAA map of the starting model.
The final model of TnC is shown in black for Figures 3.3, 3.4 and 3.5. Due to the poor quality of the starting model and the SIR phases, this map does not show many features of the final model. This Figure and Figures 3.4 and 3.5 were drawn using the program O (Jones, Zou, Cowan & Kjeldgaard, 1991).

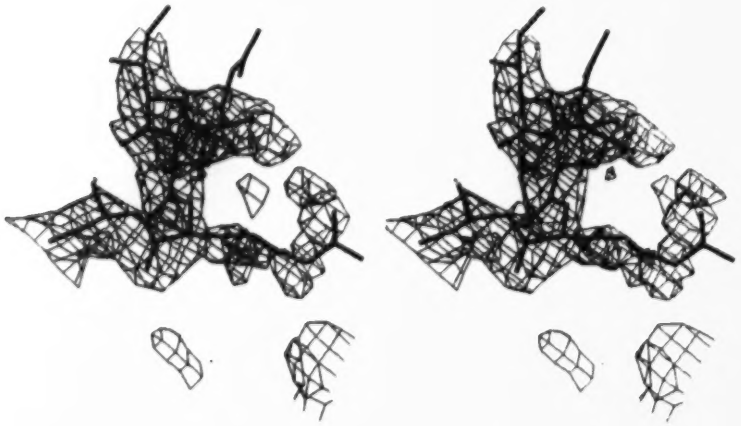


Figure 3.4: Combined phase SIGMAA map of the vector residual model

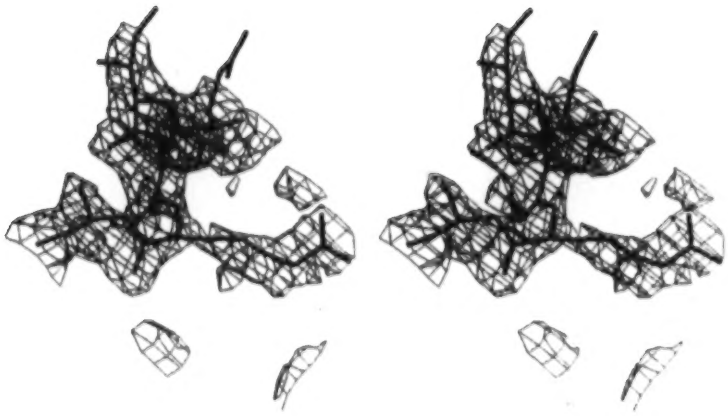


Figure 3.5: Combined phase SIGMAA map of the MLHL model

Chapter 4

Conclusions

The application of the MLF1, MLF2, and MLHL target functions to the test cases described above has yielded promising results. In the case where no prior phase information is available, the MLF1 and MLF2 targets outperform the least-squares function. As well, in the TnC test case, MLHL performed significantly better than any other target function as indicated by clearer electron density and improved phase quality. Furthermore, the difference between the working and free R factors is small in MLHL because of the inclusion of prior phase information as observations in refinement.

Although all of the test cases mentioned involved a conjugate gradients minimization scheme, the maximum likelihood target functions are in no way limited to local minimization methods. Tests have shown that the combination of a maximum likelihood target function lacking prior phase information and simulated annealing optimization parameterized in torsion angle space (Rice & Brünger, 1994) further enhances refinement (Adams, Pannu, Read & Brünger, 1997) and allows for the refinement of structures not possible by least-squares, or either method by itself. As well, recent tests have shown that the combination of torsion angle molecular dynam-

ics with the MLHL function further pushes the limits of refinement (Adams, Pannu, Read & Brünger, in preparation).

4.1 Future Developments

While the current implementations of maximum likelihood refinement already provide significant benefits, a number of improvements can be envisioned. First, the algorithm for the estimation of σ_A does not take into account measurement errors. The likelihood functions MLF1 or MLF2 can be used to compute σ_A values that take into account measurement errors. Furthermore, the likelihood function MLHL can compute σ_A values that consider prior phase information. These modified likelihood functions will be implemented in the SIGMAA algorithm in the future. As is clear from the variance term in the Gaussian approximation MLF1, observational error has little influence on the likelihood function unless the model is quite accurate. Nonetheless, it will become significant at the end of refinement and a proper treatment will be important to obtain an optimal final model.

Arbitrary relative weights between diffraction and geometry terms should not be required, in principle, if each is introduced to maximum likelihood through the appropriate probability distributions. However, some overweighting of the diffraction terms, relative to the theoretical value, is needed to achieve convergence. This may be necessary in part because the inevitable overfitting of the diffraction amplitudes alters the distribution $P(F; F_c)$. In various tests, the comparison of gradients has led to weights that are increased by factors between 4 and 50, with higher weights being required for less refined models at lower resolution. Further tests will be required to decide whether these relative weights are optimal.

The maximum likelihood approach allows one to include, in a sensible way, any

combination of information (Bricogne, 1993). Considerable scope for improvement exists in the simultaneous refinement of structures, for instance, native with liganded, or native with heavy atom derivatives. In such a refinement, all observations would be fit simultaneously, using models that are restrained to resemble one another to a degree required by the relationships among the measured sets of structure factors.

References

Adams, P.D., Pannu, N.S., Read, R.J. & Brünger, A.T. (1997). Cross-validated maximum likelihood enhances crystallographic simulated annealing refinement. Proc. Nat'l Acad. Sci. USA, 94, 5018–5023.

Baker, L (1992). C Mathematical Function Handbook. New York. McGraw-Hill.

Blundell, T.L. & Johnson, L.N. (1976). Protein Crystallography. London, Academic Press.

Bricogne, G (1991). A multisolution method of phase determination by combined maximization of entropy and likelihood III. Extension to powder diffraction data. Acta Cryst., A47, 803–829.

Bricogne, G (1993). Entropy maximization constrained by solvent flatness: A new method for macromolecular phase extension and map improvement. Acta Cryst., D49, 37–60.

Bricogne, G & Irwin, J *Macromolecular Refinement: Proceedings of the CCP4 Study Weekend January 1996*, edited by E. Dodson, M. Moore, A. Ralph & S. Bailey, pages 85–92, Daresbury, UK. Central Laboratory of the Research Councils.

Brünger, A.T. (1992). Free R value: a novel statistical quantity for assessing the accuracy of crystal structures. Nature, 355, 472–474.

Brünger, A.T., Karplus, M. & Petsko, G.A. (1989). Crystallographic refinement by simulated annealing: Application to crambin. Acta Cryst., A45, 50–61.

Brünger, A.T., Kuriyan, J. & Karplus, M. (1987). Crystallographic R factor refinement by molecular dynamics. Science, 235, 458–460.

Chambers, J.L. & Stroud, R.M. (1979). The accuracy of refined protein structures: Comparison of two independently refined models of bovine trypsin. Acta Cryst., B35, 1861–1874.

Cody, W.J. (1969). Rational Chebyshev approximations for the error function. Math. Comp., 23, 631–637.

Drenth, J (1994). Principles of Protein X-ray Crystallography. New York, Springer-Verlag.

Edwards, A.W.F. (1992). Likelihood. Baltimore, Johns Hopkins University Press.

French, S. & Wilson, K. (1978). On the treatment of negative intensity observations. Acta Cryst., A34, 517–525.

Fujinaga, M., Gros, P. & van Gunsteren, W.F. (1989). Testing the method of crystallographic refinement using molecular dynamics. J. Appl. Cryst., 22, 1–8.

Gradshteyn, I.S. & Ryzhik, I.M. (1980). Tables of Integrals, Series, and Products: Corrected and Enlarged Edition. San Diego, Academic Press.

Hendrickson, W.A. & Lattman, E.E (1970). Representation of the phase probability distributions for simplified combination of independent phase information. Acta Cryst., B26, 136–143.

Herzberg, O. & James, M.N. (1985). Structure of the calcium regulatory muscle protein troponin-c at 2.8 Å resolution. Nature, 313, 653–659.

James, R.W. (1962). The Optical Principles of the Diffraction of X-rays. Woodbridge, Ox Box Press.

Jones, T.A., Zou, J.-Y., Cowan, S.W. & Kjeldgaard, M. (1991). Improved methods for building protein models in electron density maps and the location of errors in these models. Acta Cryst., A47, 110–119.

Konnert, J.H. & Hendrickson, W.A (1980). A restrained-parameter thermal-factor refinement procedure. Acta Cryst., A36, 344–350.

Luke, Y.L. (1977). Algorithms for the computation of mathematical functions. New York, Academic Press.

Lunin, V.Y. & Urzhumtsev, A.G. (1984). Improvement of protein phases by coarse model modification. Acta Cryst., A40, 269–277.

Luzzati, V. (1952). Statistical treatment of errors in the determination of crystal structures. Acta Cryst., 5, 802–810.

Murshudov, G.N., Dodson, E.J. & Vagin, A.A. *Macromolecular Refinement: Proceedings of the CCP4 Study Weekend January 1996*, edited by E. Dodson, M. Moore, A. Ralph & S. Bailey, pages 93–104, Daresbury, UK. Central Laboratory of the Research Councils.

Otwinowski, Z. *Isomorphous Replacement and anomalous scattering: Proceedings of the CCP4 Study Weekend 25-26 January 1991*, edited by W. Wolf, P.R. Evans & A.G.W Leslie, pages 80–86, Daresbury, UK. Science and Engineering Research Council.

Read, R.J. (1986). Improved Fourier coefficients for maps using phases from partial structures with errors. Acta Cryst., A42, 140–149.

Read, R.J. (1990). Structure-factor probabilities for related structures. Acta Cryst., A46, 900–912.

Read, R.J. (1997). Model phases: Probabilities and bias. Methods in Enzymology. (In press).

Read, R.J. & James, M.N.G. (1988). Refined crystal structure of *Streptomyces griseus* at 1.7 Å resolution. J. Mol. Biol., 200, 523–551.

Rice, L.M. & Brünger, A.T. (1994). Torsion angle dynamics: Reduced variable conformational sampling enhances crystallographic structure refinement. Proteins: Structure, Function, and Genetics, 19, 277–290.

Schwarzenbach, D., Abrahams, S.C., Flack, H.D., Gonschorek, W., Hahn, Th., Huml, K., Marsh, R.E., Prince, E., Robertson, B.E., Rollet, J.S. & Wilson, A.J.C. (1989). Statistical descriptors in crystallography. Acta Cryst., A45, 63–75.

Slater, L.J. *Handbook of Mathematical Functions*, edited by M. Abramowitz & I.A. Stegun, pages 503–535. New York, Dover.

Stewart, J.M. & Karle, J. (1976). The calculation of ϵ associated with normalized structure factors, E. Acta Cryst., A32, 1005–1007.

Temme, N.M. (1983). The numerical computation of the confluent hypergeometric function U(a,b,x). Numer. Math., 41, 63–82.

Tronrud, D.E., Ten Eyck, L.F. & Matthews, B.W. (1987). An efficient general-purpose least-squares refinement program for macromolecular structures. Acta Cryst., A43, 489–501.

Wierenga, R.K., Kalk, K.H. & Hol, W.G.J. (1987). Structure determination of the glycosomal triosephosphate isomerase from *Trypanosoma brucei* at 2.4 Å resolution. J. Mol. Biol., 198, 109–121.

Wierenga, R.K., Noble, M.E.M., Vriend, G., Nauche, S. & Hol, W.G.J. (1991). Refined 1.83 Å structure of trypanosomal triosephosphate isomerase crystallized in the presence of 2.4 M ammonium sulfate. a comparison with the structure of the trypanosomal triosephosphate isomerase-glycerol-3-phosphate complex. J. Mol. Biol., 220, 995–1015.

Wilson, A.J.C. (1949). The probability distribution of X-ray intensities. Acta Cryst., 2, 318–321.

Appendix A

Deriving $\mathbf{P}(|F|; |F_c|)$

The nature of the observations made in X-ray crystallography makes it reasonable to model the four-dimensional vector $(F, F_c) = (A, B, A_c, B_c)$ as a random vector where (A, A_c) and (B, B_c) are independent bivariate Gaussians with covariance matrix

$$\epsilon \begin{pmatrix} \Sigma_N & D\Sigma_P \\ D\Sigma_P & \Sigma_P \end{pmatrix} \tag{A.1}$$

The factor D, a function of the recriprocal space vector $\mathbf{s}$, is the Fourier transform of the probability distribution of the coordinate error $(\Delta\mathbf{x})$, and intuitively represents the fraction of the calculated structure factor that is correct. As well, Σ_N is the sum of squares of the scattering factors for all N atoms in a crystal and Σ_P is the same for all P atoms in a model.

The distributions necessary for a maximum likelihood analysis of crystal structure refinement can be obtained from the joint distribution of the true structure factor F and the calculated structure factor F_c. Read (1990) has shown that the conditional distribution of F given F_c, $P(F; F_c)$ is well modelled by a two dimensional Gaussian

for acentric reflections.

$$P(F; F_c) = \frac{1}{\pi\epsilon\sigma_\Delta^2} e^{-\frac{|F - DF_c|^2}{\epsilon\sigma_\Delta^2}} \tag{A.2}$$

Multiplying the above with the Wilson (1949) distribution $P(F_c)$, we re-write the joint distribution of (F, F_c) in the following form

$$\begin{aligned} P(F, F_c) &= P(F; F_c) \times P(F_c) \\ &= \frac{1}{\pi\epsilon\sigma_\Delta^2} e^{-\frac{|F - DF_c|^2}{\epsilon\sigma_\Delta^2}} \times \frac{1}{\pi\epsilon\Sigma_P} e^{-\frac{|F_c|^2}{\epsilon\Sigma_P}} \\ &= \frac{1}{\pi^2\epsilon^2\sigma_\Delta^2\Sigma_P} e^{-\frac{|F - DF_c|^2}{\epsilon\sigma_\Delta^2} - \frac{|F_c|^2}{\epsilon\Sigma_P}} \end{aligned} \tag{A.3}$$

where $\sigma_\Delta^2 = \Sigma_N - D^2\Sigma_P$. The joint distribution of $|F|, \alpha, |F_c|$ and α_c can be obtained from expression (A.3) via a variable transformation.

$$\begin{aligned} P(|F|, \alpha, |F_c|, \alpha_c) &= |F||F_c| \times P\{F(|F|, \alpha), F_c(|F_c|, \alpha_c)\} \\ &= \frac{|F||F_c|}{\pi^2\epsilon^2\sigma_\Delta^2\Sigma_P} e^{-\frac{|F|^2 + D^2|F_c|^2 - 2D|F||F_c|\cos(\alpha - \alpha_c)}{\epsilon\sigma_\Delta^2} - \frac{|F_c|^2}{\epsilon\Sigma_P}} \end{aligned} \tag{A.4}$$

In the above expression $|F||F_c|$ is the Jacobian of the transformation.

For all the likelihood functions derived here, the distribution $P(|F|, \Delta\alpha; |F_c|, \alpha_c)$ is needed, where $\Delta\alpha$ is the difference between the true and calculated phase.

$$\begin{aligned} P(|F|, \Delta\alpha; |F_c|, \alpha_c) &= \frac{P(|F|, \Delta\alpha, |F_c|, \alpha_c)}{P(|F_c|, \alpha_c)} \\ &= \frac{P\{|F|, \alpha(\Delta\alpha, \alpha_c), |F_c|, \alpha_c\}}{P(|F_c|, \alpha_c)} \\ &= \frac{|F|}{\pi\epsilon\sigma_\Delta^2} e^{\frac{-|F|^2 - D^2|F_c|^2 + 2|F|D|F_c|\cos(\Delta\alpha)}{\epsilon\sigma_\Delta^2}} \end{aligned} \tag{A.5}$$

For the likelihood functions lacking phase information, the unknown phase error is

integrated out.

$$\begin{aligned} P(|F|;|F_c|,\alpha_c) &= \int_0^{2\pi} P(|F|,\Delta\alpha;|F_c|,\alpha_c)d\Delta\alpha \\ &= \frac{2|F|}{\epsilon\sigma_\Delta^2} e^{-\frac{|F|^2+D^2|F_c|^2}{\epsilon\sigma_\Delta^2}} I_0\left(\frac{2|F|D|F_c|}{\epsilon\sigma_\Delta^2}\right) \end{aligned} \tag{A.6}$$

The distribution obtained for $P(|F|;|F_c|,\alpha_c)$ is known as the Rice Distribution or a Non-Central χ^2 Distribution. Since this distribution does not depend on the calculated phase (α_c), the distribution is commonly denoted $P(|F|;|F_c|)$.

Similar equations can be derived for the centric case, and these expressions are given below.

$$P(|F|,\Delta\alpha;|F_c|,\alpha_c) = \frac{1}{\sqrt{2\pi\epsilon\sigma_\Delta^2}} e^{\frac{-|F|^2-D^2|F_c|^2+2|F|D|F_c|\cos(\Delta\alpha)}{2\epsilon\sigma_\Delta^2}} \tag{A.7}$$

$$P(|F|;|F_c|) = \sqrt{\frac{2}{\pi\epsilon\sigma_\Delta^2}} e^{-\frac{|F|^2+D^2|F_c|^2}{2\epsilon\sigma_\Delta^2}} \cosh\left(\frac{|F|D|F_c|}{\epsilon\sigma_\Delta^2}\right) \tag{A.8}$$

Appendix B

Implementation of MLF2

The distribution $P(J_o; J_c)$ is attained by multiplying $P(J; J_c)$ with a Gaussian probability of measurement errors ($P(J_o; J)$) with standard deviation σ_j, and integrating over the true structure factor amplitude squared, J. The distribution $P(J; J_c)$ is obtained via a variable transformation of the distribution $P(|F_o|; |F_c|)$ for acentric and centric reflections respectively.

$$P_a(J; J_c) = \frac{1}{\epsilon\sigma_\Delta^2} e^{-\frac{J+D^2J_c}{\epsilon\sigma_\Delta^2}} I_0\left(\frac{2D\sqrt{JJ_c}}{\epsilon\sigma_\Delta^2}\right) \tag{B.1}$$

$$P_c(J; J_c) = \sqrt{\frac{1}{2\pi\epsilon\sigma_\Delta^2 J}} e^{-\frac{J+D^2J_c}{2\epsilon\sigma_\Delta^2}} \cosh\left(\frac{D\sqrt{JJ_c}}{\epsilon\sigma_\Delta^2}\right) \tag{B.2}$$

The joint probability $P(J, J_o; J_c)$ is the product of the probability of the observation error and the probability of the true intensity given the calculated intensity. The desired distribution, $P(J_o; J_c)$ is the integral over J of the joint probability.

$$P(J_o; J_c) = \int_0^\infty P(J, J_o; J_c) dJ = \int_0^\infty P(J_o; J) \times P(J; J_c) dJ \tag{B.3}$$

For acentric reflections,

$$P_a(J_o;J_c) = \int_0^\infty \frac{1}{\sqrt{2\pi}\sigma_j\epsilon\sigma_\Delta^2} e^{-\frac{(J-J_o)^2}{2\sigma_j^2} - \frac{J+D^2J_c}{\epsilon\sigma_\Delta^2}} I_0\left(\frac{2D\sqrt{JJ_c}}{\epsilon\sigma_\Delta^2}\right)dJ \tag{B.4}$$

$$= \frac{1}{\sqrt{2\pi}\sigma_j\epsilon\sigma_\Delta^2} e^{-\frac{J_o^2}{2\sigma_j^2} - \frac{D^2J_c}{\epsilon\sigma_\Delta^2}} \int_0^\infty e^{-\frac{J^2}{2\sigma_j^2} - J\left(\frac{\sigma_j^2 - J_o\epsilon\sigma_\Delta^2}{\epsilon\sigma_\Delta^2\sigma_j^2}\right)} I_0\left(\frac{2D\sqrt{JJ_c}}{\epsilon\sigma_\Delta^2}\right)dJ \tag{B.5}$$

Expanding the modified Bessel function into a power series ($I_0(x) = \sum_{n=0}^{\infty} \frac{(\frac{x}{2})^{2n}}{(n!)^2}$), and interchanging integration and summation gives the following:

$$= \frac{1}{\sqrt{2\pi}\sigma_j\epsilon\sigma_\Delta^2} e^{-\frac{J_o^2}{2\sigma_j^2} - \frac{D^2J_c}{\epsilon\sigma_\Delta^2}} \sum_{n=0}^{\infty} \left(\frac{D^2J_c}{\epsilon^2\sigma_\Delta^4}\right)^n \frac{1}{(n!)^2} \int_0^\infty e^{-\frac{J^2}{2\sigma_j^2} - J\left(\frac{\sigma_j^2 - J_o\epsilon\sigma_\Delta^2}{\epsilon\sigma_\Delta^2\sigma_j^2}\right)} J^n dJ \tag{B.6}$$

There exists an antiderivative for this expression (Gradshteyn & Ryzhik. 1980).

$$= \frac{1}{\sqrt{2\pi}\epsilon\sigma_\Delta^2} e^{-\frac{J_o^2}{2\sigma_j^2} - \frac{D^2J_c}{\epsilon\sigma_\Delta^2}} \sum_{n=0}^{\infty} \left(\frac{D^2J_c\sigma_j}{\epsilon^2\sigma_\Delta^4}\right)^n \frac{1}{n!} e^{\frac{(\sigma_j^2 - J_o\epsilon\sigma_\Delta^2)^2}{4\epsilon^2\sigma_\Delta^4\sigma_j^2}} \cdot D_{-n-1}\left(\frac{\sigma_j^2 - J_o\epsilon\sigma_\Delta^2}{\epsilon\sigma_\Delta^2\sigma_j}\right) \tag{B.7}$$

The function $D_{-n-1}(x)$ is a parabolic cylinder function. Now, for centric reflections

$$P_c(J_o;J_c) = \frac{1}{2\pi\sqrt{\epsilon}\sigma_\Delta\sigma_j} \int_0^\infty \frac{1}{\sqrt{J}} e^{-\frac{J+D^2J_c}{2\epsilon\sigma_\Delta^2} - \frac{(J-J_o)^2}{2\sigma_j^2}} \cosh\left(\frac{D\sqrt{JJ_c}}{\epsilon\sigma_\Delta^2}\right)dJ \tag{B.8}$$

$$= \frac{1}{2\pi\sqrt{\epsilon}\sigma_\Delta\sigma_j} e^{-\frac{J_o^2}{2\sigma_j^2} - \frac{D^2J_c}{2\epsilon\sigma_\Delta^2}} \int_0^\infty \frac{1}{\sqrt{J}} e^{-\frac{J^2}{2\sigma_j^2} - J\frac{\sigma_j^2 - 2J_o\epsilon\sigma_\Delta^2}{2\epsilon\sigma_\Delta^2\sigma_j^2}} \cosh\left(\frac{D\sqrt{JJ_c}}{\epsilon\sigma_\Delta^2}\right)dJ \tag{B.9}$$

Expanding the Cosine hyperbolic function into a power series ($\cosh(x) = \sum_{n=0}^{\infty} \frac{x^{2n}}{(2n)!}$), and interchanging summation and integration gives

$$= \frac{1}{2\pi\sqrt{\epsilon}\sigma_\Delta\sigma_j} e^{-\frac{J_o^2}{2\sigma_j^2} - \frac{D^2 J_c}{2\epsilon\sigma_\Delta^2}} \sum_{n=0}^{\infty} \left(\frac{D^2 J_c}{\epsilon^2\sigma_\Delta^4}\right)^n \frac{1}{(2n)!} \int_0^{\infty} J^{n-\frac{1}{2}} e^{-\frac{J^2}{2\sigma_j^2} - J\frac{\sigma_j^2 - 2J_o\epsilon\sigma_\Delta^2}{2\epsilon\sigma_\Delta^2\sigma_j^2}} dJ \qquad \text{(B.10)}$$

The analytic solution for this expression is

$$= \frac{1}{2\sqrt{\pi\sigma_j\epsilon}\sigma_\Delta} e^{-\frac{J_o^2}{2\sigma_j^2} - \frac{D^2 J_c}{2\epsilon\sigma_\Delta^2}} \sum_{n=0}^{\infty} \left(\frac{D^2 J_c\sigma_j}{2\epsilon^2\sigma_\Delta^4}\right)^n \frac{1}{(2n)!!} e^{\frac{(\sigma_j^2 - 2J_o\epsilon\sigma_\Delta^2)^2}{16\epsilon^2\sigma_\Delta^4\sigma_j^2}} \mathrm{D}_{-n-\frac{1}{2}}\left(\frac{\sigma_j^2 - 2J_o\epsilon\sigma_\Delta^2}{2\epsilon\sigma_\Delta^2\sigma_j}\right) \qquad \text{(B.11)}$$

The elimination of constant terms from the above two expressions leads to the functions implemented in CNS, TNT and X-PLOR.

$$\begin{aligned}\mathcal{L} = &\sum_{hkl} \log(\epsilon\sigma_\Delta^2) + \frac{D^2 J_c}{\epsilon\sigma_\Delta^2} - \\ &\log\left(\sum_{n=0}^{\infty} \left(\frac{D^2 J_c\sigma_j}{\epsilon^2\sigma_\Delta^4}\right)^n \frac{1}{n!} e^{\frac{(\sigma_j^2 - J_o\epsilon\sigma_\Delta^2)^2}{4\epsilon^2\sigma_\Delta^4\sigma_j^2}} \mathrm{D}_{-n-1}\left(\frac{\sigma_j^2 - J_o\epsilon\sigma_\Delta^2}{\epsilon\sigma_\Delta^2\sigma_j}\right)\right)\end{aligned} \qquad \text{(B.12)}$$

for acentric reflections, and for centric reflections,

$$\begin{aligned}\mathcal{L} = &\sum_{hkl} \frac{1}{2}\log(\epsilon\sigma_\Delta^2) + \frac{D^2 J_c}{2\epsilon\sigma_\Delta^2} - \\ &\log\left(\sum_{n=0}^{\infty} \left(\frac{D^2 J_c\sigma_j}{2\epsilon^2\sigma_\Delta^4}\right)^n \frac{1}{(2n)!!} e^{\frac{(\sigma_j^2 - 2J_o\epsilon\sigma_\Delta^2)^2}{16\epsilon^2\sigma_\Delta^4\sigma_j^2}} \mathrm{D}_{-n-\frac{1}{2}}\left(\frac{\sigma_j^2 - 2J_o\epsilon\sigma_\Delta^2}{2\epsilon\sigma_\Delta^2\sigma_j}\right)\right)\end{aligned} \qquad \text{(B.13)}$$

The numerical algorithm employed to evaluate the parabolic cylinder functions can be divided into two different possibilities: one case is when the argument of the parabolic cylinder function is non-positive, and the other is when it is positive. In both cases, the algorithm developed relies on evaluating the function $\mathrm{D}_{-\nu}(x)$ for two particular values of ν, and using recursion relations to calculate the special function for the

other values of ν necessary for the series to converge.

In the first case, when the argument of the parabolic cylinder function is non-positive, the special function becomes large as $x \to -\infty$. Thus to ensure convergence of the series, $e^{-\frac{x^2}{4}} D_{-\nu}(x)$ is evaluated, and then $\frac{x^2}{2}$ is added to the likelihood function. The algorithm utilises the relationship with the complement of the error function (erfc(x)) in the acentric case:

$$D_{-n-1}(x) = \sqrt{\frac{\pi}{2}} \frac{(-1)^n}{n!} e^{-\frac{1}{4}x^2} \frac{d^n}{dx^n} (e^{\frac{1}{2}x^2} \mathrm{erfc}(\frac{x}{\sqrt{2}})) \tag{B.14}$$

where the complementary error function is defined as

$$\mathrm{erfc}(x) = \frac{2}{\sqrt{\pi}} \int_x^\infty e^{-\eta^2} d\eta \tag{B.15}$$

$$= 1 - \mathrm{erf}(x) \tag{B.16}$$

Thus, for $n = 0, 1$ equation (A.14) implies

$$e^{-\frac{x^2}{4}} D_{-1}(x) = \sqrt{\frac{\pi}{2}} \mathrm{erfc}(\frac{x}{\sqrt{2}}) \tag{B.17}$$

$$e^{-\frac{x^2}{4}} D_{-2}(x) = e^{-\frac{x^2}{2}} - x\sqrt{\frac{\pi}{2}} \mathrm{erfc}(\frac{x}{\sqrt{2}}) \tag{B.18}$$

The code for the numerical evaluation of $\mathrm{erfc}(x)$ in MLF2 was written by Cody (1969). The following recursion relation is used to calculate higher order parabolic cylinder

functions.

$$D_{-n-1}(x) = \frac{1}{n}(D_{-n+1}(x) - xD_{-n}(x)) \tag{B.19}$$

Note that both sides of the equation can be multiplied by $e^{-\frac{x^2}{4}}$ to give a recursion relation involving $e^{-\frac{x^2}{4}}D_{-n-1}(x)$.

In the centric case, when the argument of the parabolic cylinder function is non-positive, the first two terms ($n = 0, 1$) are evaluated via the relationship with the confluent hypergeometric function $\Phi(a, b, x)$:

$$\begin{aligned} e^{-\frac{x^2}{4}}D_{-n-\frac{1}{2}}(x) = &\frac{\sqrt{\pi}}{2^{\frac{n}{2}+\frac{1}{4}}}\left(\frac{1}{\Gamma(\frac{n}{2}+\frac{3}{4})}\Phi(\frac{1}{4}-\frac{n}{2},\frac{1}{2},-\frac{x^2}{2}) - \right. \\ &\left.\frac{\sqrt{2}x}{\Gamma(\frac{n}{2}+\frac{1}{4})}\Phi(\frac{3}{4}-\frac{n}{2},\frac{3}{2},-\frac{x^2}{2})\right) \end{aligned} \tag{B.20}$$

The algorithm for the numerical evaluation of $\Phi(a, b, -x)$ was adopted from Slater (1965), Luke (1977), and Baker (1992). A recursion relation similar to equation (A.19) can be used to attain higher order terms in the centric case.

In the case where the argument of the parabolic cylinder function is positive, both acentric and centric likelihood functions can be calculated using the relationship of the parabolic cylinder function with the confluent hypergeometric function $\Psi(a, b, x)$, also denoted by $U(a, b, x)$. Since $\Psi(a, b, x)$ remains bounded as x becomes large, $e^{\frac{x^2}{4}}D_{-\nu}(x)$ is evaluated.

$$e^{\frac{x^2}{4}}D_{-\nu}(x) = \frac{1}{2^{\frac{\nu}{2}}}\Psi(\frac{\nu}{2},\frac{1}{2},\frac{x^2}{2}) \tag{B.21}$$

If the first two terms ($n = 0, 1$) are evaluated using equation (A.33), and higher order terms are evaluated using equation (A.21), catastrophic cancellation occurs during the determination of higher order terms. Therefore, first $D_{-\nu}(x)$ is evaluated using

equation (A.23) for $-\nu = \lambda + 1, \lambda$, where λ is large enough to ensure convergence. Then the terms $-\nu = \lambda - 1, \lambda - 2, ..., 0$ are evaluated using a rearrangement of equation (41):

$$D_{-\nu}(x) = \nu D_{-\nu-1}(x) + x D_{-\nu-2}(x) \tag{B.22}$$

The numerical evaluation of $\Psi(a, b, x)$ in MLF2 was adopted from Temme (1983).

Note that as $(\frac{D^2 J_c \sigma_j}{\epsilon^2 \sigma_\Delta^4})$ increases, the infinite summations in expressions (A.12) and (A.13) need more terms to converge, and it is possible that the numerical values exceed machine precision before convergence occurs. We have recently derived an asymptotic equation valid for large values of $(\frac{D^2 J_c \sigma_j}{\epsilon^2 \sigma_\Delta^4})$ for acentric reflections. Such asymptotic expressions will compute the likelihood function more efficiently for large parameters and avoid potential overflow. In the two test cases discussed, however, overflow was not a problem. Nonetheless, in order to compute the likelihood function more efficiently for large parameters and avoid potential overflow, the equation derived will be implemented. In the centric case, if overflow occurs, either the MLF1 target for centric reflections can be used, or an exact probability density for the observed structure factor amplitude given the calculated amplitude (assuming a Gaussian observational error in structure factor amplitudes) that we have derived can be implemented and used.

Appendix C

Series representation of MLHL

A series representation for the MLHL function can be found by following the same derivation outlined by Hendrickson and Lattman (1970). In deriving an analytical solution for the determination of the best phases, Hendrickson and Lattman obtained a solution to the following integral:

$$\int_0^{2\pi} \exp\{A_{hl}\cos(\alpha) + B_{hl}\sin(\alpha) + C_{hl}\cos(2\alpha) + D_{hl}\sin(2\alpha)\}\, d\alpha = 2\pi\left\{\mathrm{I}_0(S)\mathrm{I}_0(T) + 2\sum_{n=0}^{\infty}\mathrm{I}_{2n}(S)\mathrm{I}_n(T)\cos(n(2\sigma-\tau))\right\} \tag{C.1}$$

- $S = \sqrt{A_{hl}^2 + B_{hl}^2}$
- $T = \sqrt{C_{hl}^2 + D_{hl}^2}$
- $\tan(\sigma) = -\frac{B_{hl}}{A_{hl}}$
- $\tan(\tau) = -\frac{D_{hl}}{C_{hl}}$

The required integral for the MLHL function can be written in the above form, giving the following solution:

$$P(|F|;|F_c|,\alpha_c) = \frac{2N}{\epsilon\sigma_\Delta^2}\exp\left(\frac{-|F|^2 - D^2|F_c|^2}{\epsilon\sigma_\Delta^2}\right) \times \left\{ \mathrm{I}_0(S')\mathrm{I}_0(T) + 2\sum_{n=0}^{\infty} \mathrm{I}_{2n}(S')\mathrm{I}_n(T)\cos(n(2\sigma' - \tau)) \right\} \tag{C.2}$$

where

- $S' = \sqrt{A_o^2 + B_o^2}$
- $\tan(\sigma') = -\frac{B_o}{A_o}$
- $A_o = A_{hl} + \frac{2|F|D|F_c|\cos(\alpha_c)}{\epsilon\sigma_\Delta^2}$
- $B_o = B_{hl} + \frac{2|F|D|F_c|\sin(\alpha_c)}{\epsilon\sigma_\Delta^2}$

The series has a cosine term, and consequently can take both positive and negative values. Because of rounding off errors, for particular arguments of the function, the series representation can result with an undefined negative probability. Thus, we have chosen to evaluate the function numerically in the general case of non-zero Hendrickson-Lattman coefficients.

Appendix D

Phased likelihood with measurement errors

In order to derive a likelihood function incorporating prior phase information that include the effect of measurement error of the native structure factor amplitude, the joint probability distribution, $P(|F|, \Delta\alpha, \alpha; |F_c|, \alpha_c)$ must be multiplied by a probability distribution of the observed structure factor amplitude given the true structure factor amplitude, $P(|F_o|; |F|)$. The resulting expression is the joint probability distribution $P(|F_o|, |F|, \Delta\alpha, \alpha; |F_c|, \alpha_c)$. The required distribution is obtained by integrating out the true structure factor amplitude and phase:

$$P(|F_o|; |F_c|, \alpha_c) = \int_0^{2\pi} \int_0^{\infty} P(|F|, \Delta\alpha, \alpha; |F_c|) \times P(|F_o|; |F|) d|F| d\alpha \qquad \text{(D.1)}$$

In this derivation, a Gaussian probability distribution of the observed structure factor amplitude given the true structure factor amplitude will be assumed. As well, only acentric reflections will be considered here, but similar equations can be derived for

the centric case. The required integral for the acentric case is the following.

$$P(|F_o|;|F_c|,\alpha_c) = \frac{1}{\sqrt{2\pi^3}\sigma_F\epsilon\sigma_\Delta^2}\int_0^{2\pi} P(\alpha)\int_0^{\infty}|F|\times \exp\left\{-|F|^2\left(\frac{1}{2\sigma_F^2}+\frac{1}{\epsilon\sigma_\Delta^2}\right)+|F|\left(\frac{F_o}{\sigma_F^2}+\frac{2D|F_c|\cos(\Delta\alpha)}{\epsilon\sigma_\Delta^2}\right)\right\} d|F|d\alpha \tag{D.2}$$

The true structure factor amplitude can be integrated out of this expression (Gradshteyn & Ryzhik, 1980), leaving only a numerical integration of the true phase.

$$P(|F_o|;|F_c|,\alpha_c) = \frac{\sigma_F}{\sqrt{2\pi^3}(\sigma_F^2+\epsilon\sigma_\Delta^2)}\exp\left(-\frac{|F_o|^2}{2\sigma_F^2}-\frac{D^2|F_c|^2}{\epsilon\sigma_\Delta^2}\right) \int_0^{2\pi} P(\alpha)\left\{1+\nu\sqrt{\pi}\exp(\nu^2)\mathrm{erfc}(-\nu)\right\} d\alpha \tag{D.3}$$

where $\nu = \frac{|F_o|\epsilon\sigma_\Delta^2+2D|F_c|\cos(\alpha-\alpha_c)\sigma_F^2}{\sigma_F}\sqrt{\frac{\epsilon\sigma_\Delta^2+2\sigma_F^2}{2\epsilon\sigma_\Delta^2}}$